VŠB – Technická univerzita Ostrava
Fakulta elektrotechniky a informatiky
Katedra informatiky

Detekce žánrů webových stránek

Web genre detection

2012

Michael Waligora

VŠB - Technická univerzita Ostrava
Fakulta elektrotechniky a informatiky
Katedra informatiky

Zadání bakalářské práce

Student: **Michael Waligora**

Studijní program: B2647 Informační a komunikační technologie

Studijní obor: 2612R025 Informatika a výpočetní technika

Téma: Detekce žánrů webových stránek
Web Genre Detection

Zásady pro vypracování:

Cílem práce je provedení průzkumu existujících přístupů, návrh a implementace vybrané nebo vlastní metody a aplikačního prostředí pro experimenty.

1. Průzkum a popis existujících přístupů.
2. Návrh a implementace vybrané nebo vlastní metody.
3. Návrh a implementace počítačové aplikace pro provádění experimentů.
4. Návrh, realizace a hodnocení experimentů.

Seznam doporučené odborné literatury:

Podle pokynů vedoucího bakalářské práce.

Formální náležitosti a rozsah bakalářské práce stanoví pokyny pro vypracování zveřejněné na webových stránkách fakulty.

Vedoucí bakalářské práce: **Mgr. Miloš Kudělka, Ph.D.**

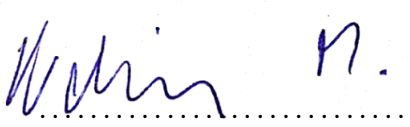
Datum zadání: 18.11.2011
Datum odevzdání: 04.05.2012

doc. Dr. Ing. Eduard Sojka
vedoucí katedry

prof. RNDr. Václav Snášel, CSc.
děkan fakulty

Prohlašuji, že jsem tuto bakalářskou práci vypracoval samostatně. Uvedl jsem všechny literární prameny a publikace, ze kterých jsem čerpal.

V Ostravě 4. května 2012



Rád bych poděkoval Mgr. Miloši Kudělkovi PhD. za cenné rady při vedení mé bakalářské práce.

Abstrakt

Hlavním cílem práce je prozkoumání existujících metod automatické detekce žánrů webových stránek, dále vybranou metodu implementovat pro detekci několika žánrů a otestovat.V práci bude popsaná stručná definice webových žánrů, čím se odlišují, definice vlastností webových stránek. Dále budou popsány předchozí prováděné experimenty a některé existující metody detekce. Aby mohla být přesnost vybrané metody ověřena, bude navržena desktopová aplikace pro stahování webových stránek, parsování webových stránek a detekci jejich žánrů. Nakonec bude přesnost detekce mojí metody porovnána s jinými metodami.

Klíčová slova: Detekce webových žánrů, Webový žánr, Parsování, Naivní bayesovský klasifikátor

Abstract

The main aim of my work is exploration of existing methods of automatic web genre detection, then I implement and test the chosen method for several genres. In my work definition of web genre, differences between each other and features of web genres will be described briefly. Then previous solved experiments and some existing methods of detection will be described. In order to check accuracy of chosen method desktop application will be designed. It will be able to downloading and parsing a web pages and detect its genre. Finally accuracy of my method will be compared with other methods.

Keywords: Detection of web genre, Web genre, Parsing, Naive Bayes classifier

Seznam použitých zkratek a symbolů

CFS	–	Correlation-based Feature Subset Selection
DOM	–	Document Object Model
FAQ	–	Frequently Asked Questions
HAP	–	Html Agility Pack
HTML	–	Hyper Text Markup Language
URL	–	Uniform Resource Locator
XHTML	–	Extensible HyperText Markup Language
XPATH	–	XML Path Language
XSLT	–	Extensible Stylesheet Language Transformations

Obsah

1 Úvod

V dnešní době internetu, který nám poskytuje téměř nekonečné množství informací, se může zdát vyhledávání informací snadné. Ale zkusme si vyhledat slovo auto, vyhledávač Google vám nabídne přes 3 miliardy stránek. V takovém množství informací se uživatel může snadno ztratit. Právě v takových situacích můžeme využít detekci žánrů webových stránek, pro upřesnění našeho dotazu. Například po zobrazení jenom zpravodajských stránek se nám počet stránek zúží o několik řádů na desítky tisíc.

Za žánr lze obecně považovat něco co má stejné vlastnosti, účel nebo formu a je jedno k jakému odvětví se to vztahuje. Filmové, divadelní nebo literární žánry zná každý. Pod slovy horor, povídka, bajka nebo drama si určitě každý dokáže představit jejich příklady a vlastnosti. Kdybych se vás zeptal uveďte mi příklady webového žánru? Každý z vás by jistě několik uvedl, ale mysleli byste pod stejným slovem stejnou stránku? A to je největší problém, že neexistuje žádná oficiální definice co to žánr je a co není. Jaké jsou jeho hranice? Kde začíná zpravodajská stránka a kde blog? Jestli firemní stránka, kde je odkaz na koupi produktu je už elektronický obchod? Žánr dokumentu snadno rozeznávají ti, kteří dokument vytvořili nebo ho používají [2]. Například starší generace bude těžko rozeznávat blog, protože nikdy neviděli internet. A naopak mladší generace ho snadno rozezná.

Cílem bakalářské práce je popsat existující přístupy detekce žánrů webových stránek. Dalším cílem je vlastní implementace existující nebo vlastní metody a její otestování.

V kapitole 2 bude popsána definice žánru, vybrané žánry pro detekci, vlastnosti stránek a existující metody detekce žánrů.

Třetí kapitola popisuje vlastní implementaci mojí aplikace. Postupně je popsán přistup k parsování dokumentu a Naivní bayesovská metoda, která byla vybrána pro detekci žánru.

Čtvrtá kapitola popisuje experimenty. Jsou zde popsány a zhodnoceny výsledky jednotlivých testů, které jsou porovnány s ostatními klasifikačními metodami.

2 Žánr a jeho klasifikace

2.1 Definice žánru

Jedna z možností jak definovat žánry je pomocí průzkumu, kdy se vytvoří sada dokumentů, které nejsou přiřazeny do žádného žánru. Každý účastník průzkumu ručně přiřadí dokument do žánru, nejčastěji přiřazovaný žánr je potom považován za žánr dokumentu. Více je tato metoda popsána zde [2].

Dobrou pomůckou pro identifikaci žánru je komunikační účel dokumentu a jeho funkční rysy. Např. jestli nám dokument slouží k vyměňování názoru, potom se jedná s vysokou pravděpodobností o diskuzi. Funkce a účel dokumentu je tedy většinou snadno rozeznatelný člověkem, ale hůře počítačem. Naopak počítač snadno rozezná jednoduché vlastnosti jako počet HTML tagů, počet znaků, počet klíčových slov atd. Bohužel neexistuje žádná definice kolik těchto jednoduchých vlastností žánr má. Možnost je extrahovat tyto vlastnosti z trénovací sady dokumentů, které by měly být z různých domén a tématických oblastí, aby se pokryl co největší rozsah dokumentů.

Problém webových dokumentů je, že není žádná šablona žánru a každý vývojář tvoří stránky jinak. Někdo používá místo tagu hlavního nadpisu h1 div, místo odstavce p tabulku apod. Kvůli tomu je obtížné najít normovaný dokument k žánru podle HTML tagů.

Ve většině předchozích pracích je žánr webové stránky definován stylem, formou a obsahem [9, 1]. Stylem se rozumí jak je dokument napsán. Jsou zde obsaženy vlastnosti jako interpunkce, velikost slov nebo délka vět. Forma nám popisuje jak je dokument zobrazen, jaký layout stránky je použit. Některé stránky mají složitější strukturu (např. eshop), jiné jsou jednodušší skládající se pouze z jednoho formuláře (např. webový vyhledávač). Obsah zahrnuje vlastnosti jako počet výskytů specifických slov, například na stránce se žánrem blog je velká pravděpodobnost výskytu tohoto slova.

Další důležitou charakteristikou žánru je jeho nezávislost na tématu, tzn. množina stránek stejného žánru zahrnuje několik témat nebo jedno téma se vztahuje k několika žánrům.

V současnosti je těžké určit přesný počet webových žánrů, v rychlém vývoji na internetu neustále vznikají nové žánry. Před několika lety nikdo nevěděl, že bude existovat

žánr sociálních sítí nebo aukce. Jenom na stránkách webgenrewiki je popsáno kolem 80 existujících žánrů [9] a jistě to není konečný seznam.

Z důvodu, že žánry webových stránek se můžou lišit podle jazyka, jsem se rozhodl v mé práci detekovat pouze stránky v anglickém jazyce.

2.2 Vybrané žánry pro testování

K testování byly vybrány tyto žánry: blog, elektronický obchod, často kladené dotazy (FAQ) a diskuzní fórum.

Blog

Nejběžnější osobní blog je stránka, kde většinou přispívá jeden blogger, který vydává články (posts). Existují však i jiné typy blogů: firemní blog, tématický blog, podle typu média[1], podle typu zařízení[2]. V mé práci budou detekovány osobní a tématické blogy [14].

Charakteristika:

- funkce: možnost přidání komentáře
- struktura: seznam článků řazených chronologicky od nejnovějšího
- styl: dlouhé věty
- obsah: informace o jednom nebo více tématech

Často kladené dotazy (FAQ)

FAQ (zkratka anglického výrazu Frequently Asked Questions) je dokument, kde jsou zobrazeny časté dotazy začátečníku a jejich odpovědi. Původně internetový dokument se později rozšířil do dalších forem, FAQ dokumenty jsou nyní často součástí softwarové nápovědy [15].

Charakteristika:

- funkce: žádná funkcionalita
- struktura: seznam

[1] vlog - video blog, linklog - obsahuje odkazy, sketchblog, fotoblog
[2] moblog - blog psaný mobilním telefonem

- styl: dlouhé věty
- obsah: páry otázek a odpovědí

Elektronický obchod (Online shopping)

Elektronický obchod je forma elektronické komerce, kde zákazníci přímo kupují zboží nebo služby od prodejce.

Charakteristika:

- funkce: nákup zboží
- struktura: tabulková struktura, hierarchické rozdělení do kategorií
- styl: krátké věty, fráze
- obsah: katalogy zboží

Diskuzní fórum

Diskuzní fórum je obvykle tématicky zaměřená stránka, uživatelé si můžou vyměňovat své názory na dané téma. Kvalitu diskuze zajišťují moderátoři, kteří filtrují nevhodné příspěvky. Diskuze je členěna do vláken vztahují se k určité otázce.

Charakteristika:

- funkce: možnost založení nového vlákna, přidávat příspěvky
- struktura: vlákna jsou zobrazena v seznamu a členěna do kategorií
- styl: diskutování problému nebo otázky ve vláknu
- obsah: podle tématu diskuze

2.3 Vlastnosti

Každý dokument je charakterizován svými vlastnostmi, pomocí kterých jsme schopni dokument přiřadit do žánru. Klasické textové dokumenty jsme mohli charakterizovat pouze pomocí stylu psaní a obsahu. Webové HTML dokumenty obsahují navíc další informace v URL a HTML značích. Těchto vlastností existují stovky, ale ne všechny se dají použít pro přiřazení dokumentu do žánru. Jako výchozí množinu všech vlastností jsem použil upravenou množinu, kterou použili zde [17]:

- frekvence HTML tagů (viz. tabulka 1)
- frekvence 50 nejběžnějších slov v angličtině (viz. tabulka2)
- frekvence interpunkčních znamének - dvojtečka, středník, čárka, tečka, vykřičník, otazník, apostrof, uvozovky
- slova specifická pro žánr
- počet znaků - počet znaků čistého textu (nejsou započítávány znaky ve skriptech, atributech, komentářích...)
- průměrná délka věty

<a>	<h1>	<object>
<alt>	<h2>	
<applet>	<h3>	<p>
<b>	<h4>	<script>
 	<hr>	
	<mailto>	<table>
<dl>	<i>	<u>
	<img>	
<font>	<input>	
<form>	<noscript>	

Tabulka 1: 28 použitých HTML tagů

Nejběžnější slova v angličtině

V předchozích experimentech se ukázalo pro detekci žánru užitečné použití těchto 50 nejběžnějších slov v angličtině nashromážděných z BNC (British National Corpus).

Slova specifická pro žánr

Tyto množiny slov, které jsou tématicky nezávislé, se často vyskytují v žánrech:

- **slova specifická pro blog** - blog, comment, diary, journal, posted, archive, response;
- **slova specifická pro elektronický obchod** - basket, buy, cart, catalogue, checkout, cost, credit card, debit card, delivery, offer, order, pay, price, purchase, rebate, save, sell, shipping, shop, store, story, trolley;

the	with	are	or	her
of	he	not	an	n't
and	be	his	were	there
a	on	this	we	can
in	i	from	their	all
to	that	but	been	as
is	by	had	has	if
was	at	which	have	who
it	you	she	will	what
for	's	they	would	said

Tabulka 2: 50 nejběžnějších slov v angličtině

- **slova specifická pro FAQ** - faq, frequently asked question, answer, enquir, inquir;
- **slova specifická pro diskuzi** - board, bulletin, fori, forum, mailing, message, moderator, post, proble, story, thread, topic, user.

2.4 Funkčně motivované vlastnosti

Jiný pohled na vlastnosti podle [4] uvádí, že vlastnost dokumentu může reprezentovat tzv. facet. Pod pojmem facet si můžeme představit pohled na situaci, kde každý facet reprezentuje funkci nebo vlastnost. Facet je makro-vlastnost, tzn. každý facet se skládá z několika malých vlastností.

Tento přístup má výhodu, že jsou pro člověka snadněji pochopitelné, než jednoduché vlastnosti dokumentu uvedené výše. Díky tomu můžeme snadněji usoudit do kterého žánru dokument patří. Pro detekci blogu jsem se pokusil vytvořit tyto facety:

- jsou pod článkem komentáře
- je u názvu článku uvedeno datum vydání

Princip detekce spočíval v nalezení nejčastějšího nadpisu v dokumentu (h1 ... h4). Poté byl spočítán poměr nejčastější nadpis / počet výskytů data a nejčastější nadpis / počet výskytů slova komentář. Pokud byl výsledek v intervalu <0.4;1.6> byl facet považován za detekovaný. Tento interval byl odhadnut pouhým pozorováním dokumentů v trénovací sadě a pravděpodobně nemůže být použit u jiných dokumentů, ale pro moje experimenty se tato metoda ukázala jako dostačující.

2.5 Klasifikace

Automatická klasifikace žánru je v ideálním případě proces přiřazení webového dokumentu do žádného (pokud je dokument hodně individuálně upraven), jednoho (pokud dokument náleží do jednoho žánru) nebo více žánrů (když dokument obsahuje několik žánrů nebo je hybrid)[4]. V mé práci se budu zabývat pouze klasifikací jednoho žánru.

Běžné metody klasifikace jsou založeny na algoritmu učení s učitelem [6]. Učení s učitelem je metodou strojového učení (SU), která se zabývá algoritmy umožňující strojům proces učení.

Klasifikace žánru pomocí učení s učitelem spočívá ve vytvoření trénovacích dat (množina webových dokumentů), které jsou přiřazeny do určitého žánru. Algoritmus poté extrahuje vlastnosti těchto dokumentů. Nový, dosud do žádného žánru nepřiřazený dokument klasifikátor přiřadí do nejvíce podobného žánru. Cílem je optimalizace výběru vhodných vlastností, abychom dosáhli co nejpřesnějších výsledků. Postupným přidáváním vlastností můžeme jak zvyšovat přesnost klasifikátoru, tak i snižovat. Příčinou tohoto problému je naučení se závislostí, které ve skutečnosti neexistují. Tento problém se nazývá přeučení [5].

Problémem u detekce žánrů webových dokumentů je, že obsahují příliš mnoho neužitečných informací pro detekci. Dnes používané značkovací jazyky HTML 4 a XHTML postrádají sémantické prvky. Proto vzniká např. problém, kde hledat relevantní informace o průměrné délce věty. Typická stránka se skládá z jednoho hlavního panelu (zde je umístěn obsah dokumentu) a z několika postranních panelů (zde jsou umístěny často navigační menu, reklamní banery atd.). Relevantní informace by se měli hledat v hlavní části. Ale v současnosti kdy je stránka rozdělena do divů označených libovolným id nebo třídou není snadné najít ten požadovaný div. Z tohoto důvodu můžeme při parsování dokumentu získat data které se nijak nevážou k žánru a mají negativní vliv na klasifikaci. Tyto data se nazývají šum. HTML 5 by mohlo přinést zlepšení, protože zavádí nové sémantické značky - aside (boční panel), nav (umístění navigace), article (umístění článku, komentáře), header (hlavička dokumentu), footer(patička dokumentu) atd.

2.6 Metody

Některé používané metody detekce žánrů v minulosti [1]:

- TF*IDF
- K-nejbližší soused
- Naivní bayesovský klasifikátor
- Bayesovské sítě
- Deduktivní metoda
- Support Vector Machines
- LogitBoost
- Rozhodovací stromy

TF*IDF

První metodou je TF*IDF, která byla použita např. zde [7]. TF (term frequency) znamená počet výskytu termínu v dokumentu - čím více výskytů, tím větší váha termínu. IDF (inverse document frequency) značí zda je termín běžný v trénovací sadě - čím více dokumentů obsahující termín, tím menší váha termínu. $IDF = \log \frac{N}{DF}$, kde DF je počet dokumentů obsahující daný termín a N je celkový počet dokumentů. Termín může být slovo, nebo sousloví. Po zparsování dokumentu a spočítání vah jednotlivých termínů můžeme využít několik algoritmů ke klasifikaci: Naivní bayesovský klasifikátor, Support Vector Machines (SVM) nebo algoritmus K-nejbližší soused.

Algoritmus K-nejbližší soused (K-Nearest Neighbour)

Je algoritmus založený na paměti, který nepotřebuje trénovací fázi. Příklady jsou reprezentovány vektory o n komponentách, kde n je počet atributů (vlastností). Ke klasifikaci nového dokumentu algoritmus porovná dokument se všemi případy trénovací množiny a vypočte vzdálenost mezi nimi. Potom je dokument přiřazen do většinové třídy pro K nejpodobnějších případů. Podobnost sousedů je reprezentována většinou euklidovskou metrikou. Například pokud 3 ze 4 sousedů budou ze stejného žánru, potom daný dokument přiřadíme do stejného žánru. Pokud jsou všichni sousedi odlišní, přiřadí se dokument k nejbližšímu dokumentu [8].

Naivní bayesovský klasifikátor

Naivní bayesovský klasifikátor je jednoduchý pravděpodobnostní algoritmus, který určí s jakou pravděpodobností dokument patří do daného žánru. Výhodou tohoto algoritmu je velmi rychlé učení a odolnost vůči nevýznamným parametrům. Jelikož váhy jsou stejně velké pro všechny parametry, přesnost algoritmu může být snížena použitím příliš mnoha parametrů. Přes svoji jednoduchost algoritmus prokazuje velkou přesnost v mnoha oblastech. Tento algoritmus byl vybrán pro implementaci a bude podrobněji popsán v kapitole 3.3 [10].

LogitBoost

LogitBoost patří do kategorie boostovacích algoritmů, které představují metodu strojového učení pomocí tzv. meta-algoritmu (metody jak nejlépe učit klasifikátory). Principem boostingu je vytvořit silný klasifikátor pomocí iteračního učení slabých klasifikátorů. Při iterování slabých klasifikátorů se upravují váhy vlastností tak aby bylo dosaženo co nejlepších výsledků [12].

Deduktivní metoda

Deduktivní metoda (Inferential model) spoléhá na funkčně motivované vlastnosti, metoda implementuje zjednodušenou formu bayesovského teorému nazývanou odds-likelihood nebo subjektivní bayesovskou metodu. Hlavním rozdílem oproti klasického bayesovskému klasifikátoru je, že váhy jednotlivých parametrů jsou různé. Proto je podstatný správný výpočet vah k přesné klasifikaci. Pro konečnou klasifikaci se používá jednoduché rozhodovací IF-THEN pravidlo [4].

Support Vector Machines (SVM)

SVM je binární lineární klasifikátor, kde jsou třídy příkladů z trénovacích dat reprezentovány jako body v prostoru. Jednotlivé třídy jsou odděleny nadrovinou[3] tak, aby mezi nimi byla co největší vzdálenost. Ke klasifikaci nového příkladu se nový příklad zobrazí ve stejném prostotu jako trénovací množina a je klasifikován podle toho v jakém prostoru se zobrazí [13].

[3]Nadrovina pro daný prostor dimenze n je jeho podprostor dimenze n-1.

Porovnání přesnosti metod

V tabulce 2.6 jsou zobrazeny orientační výsledky přesností klasifikačních metod prováděné v předchozích experimentech. K testování byly použity trénovací sady obsahující 1 400 dokumentů (pro SVM a Naivní bayes) a 2 480 dokumentů (pro odvozovací metodu) [4].

Žánr	SVM (%)	Naivní bayes (%)	Odvozovací metoda (%)
Blog	96	92	91
Eshop	88	76	83
FAQ	94.5	67	88.5
Front page	100	98	97
Seznamy	80	29	75.5
Osobní stránky	79	27	77
Vyhledávače	85	82	88
Celkem	89	67	86

Tabulka 3: Porovnání přesnosti klasifikačních metod

3 Implementace klasifikační metody

K implementaci aplikace byl použit .NET framework a jazyk C#. .NET framework jsem vybral z důvodu rychlého vývoje GUI na Windows platformě pomocí WinForm a velkého množství knihoven pro prací s HTML kódem.

3.1 Parsovaní

K parsování HTML stránek byla použita .NET knihovna pod Microsoft Public License HTML Agility Pack (HAP). HAP je agilní HTML parser který nám umožňuje:

- Načtení a uložení stránky z několika zdrojů - web, soubor.
- Funkce pro manipulaci s DOM - přidávání/odebírání uzlů, čtení hodnot uzlů, počet uzlů ...
- Prohledávání struktury HTML dokumentu pomocí technologií XPATH a XSLT (příklad na výpisu 1).

```
parsedDoc.Genre = document.DocumentNode.SelectSingleNode("//div[@id='genre_type']").
    InnerText;
//obsahuje url blog, pri ulozeni dokumentu byl pridan div s url
parsedDoc.Url = document.DocumentNode.SelectSingleNode("//div[@id='url']").InnerText;
```

Výpis 1: Příklad získání hodnoty uzlu podle hodnoty id

Pomocí HAP tedy můžeme snadno zjišťovat počet HTML tagů, hodnoty atributů nebo hodnoty tagu. Problém však nastává když chceme např. zjistit počet znaků v dokumentu, počet vět, počet termínů atd. Vlastnost InnerText třídy DocumentNode vrací řetězec obsahující také nežádoucí prvky (komentáře, skripty, vnořené html tagy atd.). Například když chceme vyhledat slovo comment, většina vývojářů si dá do kódu před div s komentářem poznámku, že se tady zobrazují komentáře (v textu je jeden výskyt comment, parser najde dva výskyty). Tento problém se pak negativně projevuje na přesnosti parseru. Pro jsem vytvořil funkci ConvertToText(string html) viz. Výpis 5, která nám odfiltruje nežádoucí obsah. Příklad kódu hledající počet výskytu klíčových slov najdete na Výpisu 2. K ukládání extrahovaných vlastností jsem vytvořil třídu DocumentInformation, která je v ukázkách kódu reprezentována objektem parsedDoc.

```
string allString = document.DocumentNode.SelectSingleNode("//body").InnerHtml.Trim();
allString = ConvertToText(allString);
//hledani faq words
matchCol = Regex.Matches(allString, @"(faq)|(frequently\sasked\squestion)|(answer)|(enquir)|(
    inquir)", RegexOptions.Compiled | RegexOptions.IgnoreCase);
parsedDoc.FaqWordCount = matchCol.Count;
```

Výpis 2: Příklad zjištění počtu klíčových slov

3.2 IF-THEN klasifikace

První algoritmus, který jsem použil byla jednoduchá IF-THEN klasifikace. Cílem je pouhé rozlišení zda daný dokument patří do žánru, nebo nepatří. Klasifikace využívá relativně málo funkčně motivovaných vlastností se stejnou váhou. Pokud dokument obsahuje danou vlastnost je k celkovému skóre přičtena 1. Dokument je pozitivně klasifikován pokud je součet skóre alespoň 60% maximálního (viz. Výpis 3). Později jsem od tohoto algoritmu upustil z důvodu problematické detekci jeho vlastností a implementoval jsem Naivní bayesovský klasifikátor viz. kapitola 3.3.

```
int blogCandidate = 0;
if (this.ContainBlogWords)
    blogCandidate++;
if (this.containComment)
    blogCandidate++;
if (this.containDates)
    blogCandidate++;
if (blogCandidate >= 2)
    return true;
else
    return false;
```

Výpis 3: Příklad IF-THEN klasifikace blogu

3.3 Naivní bayesovský klasifikátor

Naivní bayesovský klasifikátor vychází z Bayesovy věty o podmíněné pravděpodobnosti. Tato pravděpodobnostní metoda je velice často využívána ve strojovém učení ke klasifikaci. Bayesovský vztah (1) slouží pro výpočet aposteriorní pravděpodobnosti P(H|E), tedy podmíněné pravděpodobnosti, že platí hypotéza H při pozorování evidence E. Vychází z apriorní pravděpodobnosti hypotézy P(H), pravděpodobnosti výskytu evidence

P(E) a podmíněné pravděpodobnosti P(E|H), která popisuje pozorování evidence E v případě, že platí hypotéza H [10].

$$P(H|E) = \frac{P(E|H)P(H)}{P(E)} \tag{1}$$

První metodou můžeme spočítat pravděpodobnost pro jednu evidenci, pro sledování více evidencí použijeme vzorec pro Naivní bayesovský klasifikátor (2). Vycházíme z předpokladu, že jednotlivé evidence jsou nezávislé [10].

$$P(H|E_1, ..., E_K) = \frac{P(H)}{P(E_1, ..., E_K)} \prod_{k=1}^{K} P(E_K|H) \tag{2}$$

Pro výpočet pravděpodobností pro evidenci x je použito normální (nebo Gaussovo) rozdělení pravděpodobnosti (3) [11]. Z trénovacích dat získáme střední hodnotu π a rozptyl σ. Příklad vzorce v jazyku C# můžete vidět na Výpisu 4.

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{(x-\mu)^2}{2\sigma^2}} \tag{3}$$

```
private double probability(Parameter p, double x)
    {
        return (1 / (Math.Sqrt(2 * Math.PI * p.Variance))) * Math.Exp((-(x - p.Mean) *(x - p
            .Mean))/(2*p.Variance));
    }
```

Výpis 4: Zjednodušený kód výpočtu pravděpodobnost pro jeden parametr

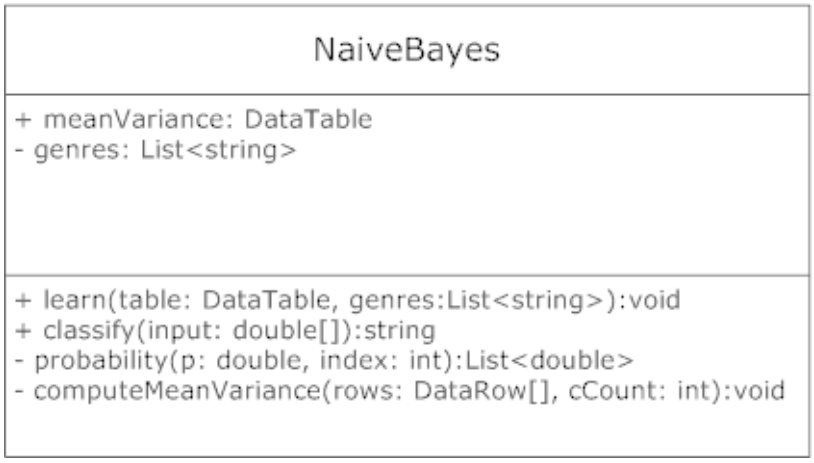
V případě, že existuje více hypotéz T a rozhodujeme, která je pro danou evidenci nejpravděpodobnější. Podle vzorce (4) vybereme hypotézu s největší aposteriorní pravděpodobností H_{MAP} z prostoru hypotéz T. Jelikož nás nezajímá přesná hodnota pravděpodobnosti ale pouze největší, můžeme ve vzorci vynechat jmenovatele, který je pro všechny hypotézy stejný.

$$P(H|E_1, ..., E_K) = argmax P(H_t) \prod_{k=1}^{K} P(E_K|H), t \in T \tag{4}$$

Algorimus

- Vybere se trénovací sada pro učení.
- Parser extrahuje vybrané vlastnosti.
- Klasifikátor se naučí - spočítá rozptyl a průměr pro každou vlastnost pro všechny žánry.
- Vybere se testovací sada, pro každý prvek v testovací sadě se spočítá pravděpodobnost pro všechny žánry, prvek je klasifikován k žánru s nejvyšší pravděpodobností.

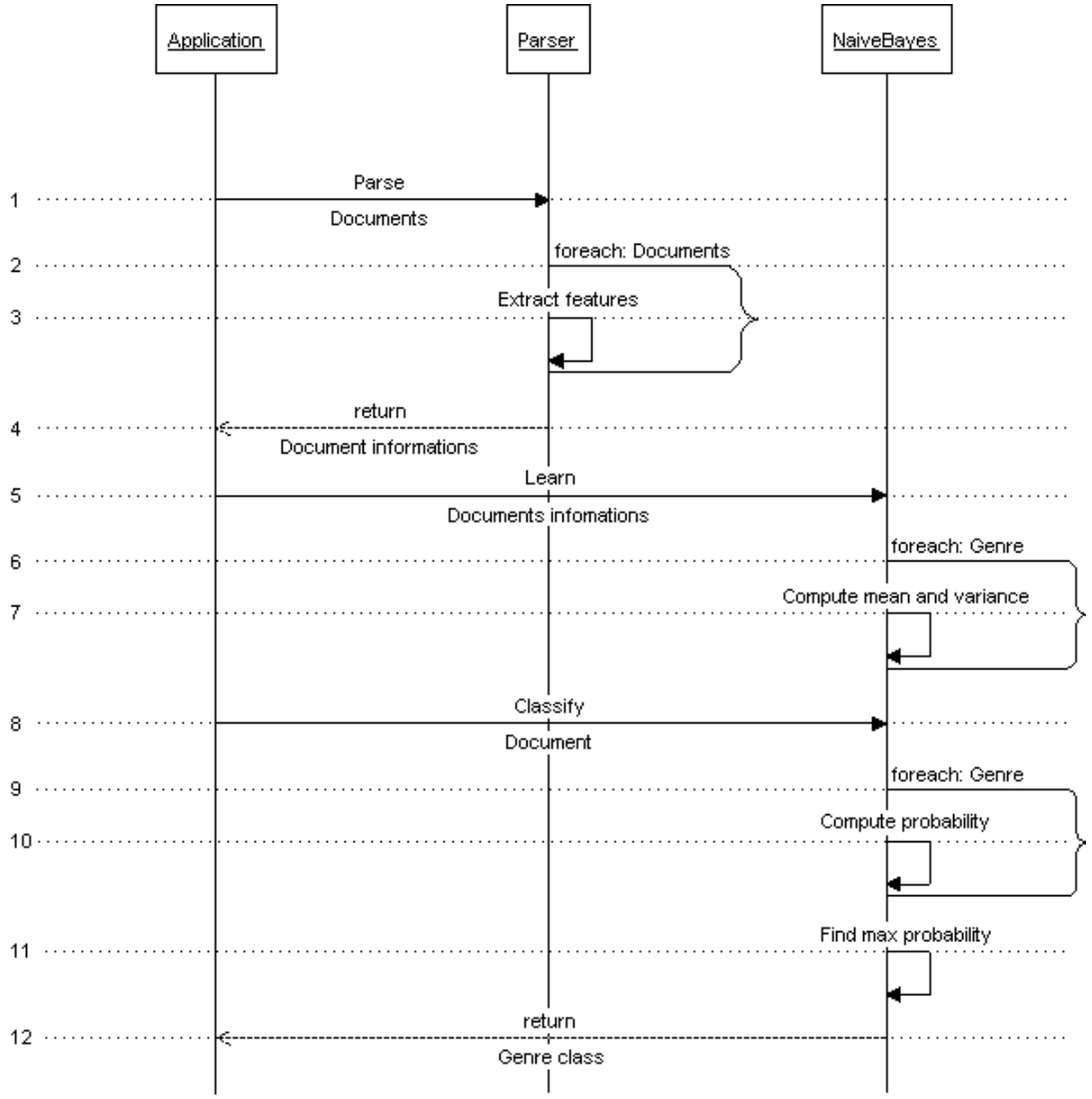
Pro aplikování Naivního bayesovského klasifikátoru jsem vytvořil třídu NaiveBayes, jejíž třídní diagram je zobrazen na obrázku 1. Tato třída není omezena pouze na detekci žánrů webových stránek, ale dá se využít pro libovolná data. Jediným omezením je, že hodnoty vlastností musí být typu double.



Obrázek 1: Třídní diagram třídy NaiveBayes

Pro používání třídy stačí znát tyto veřejné metody:

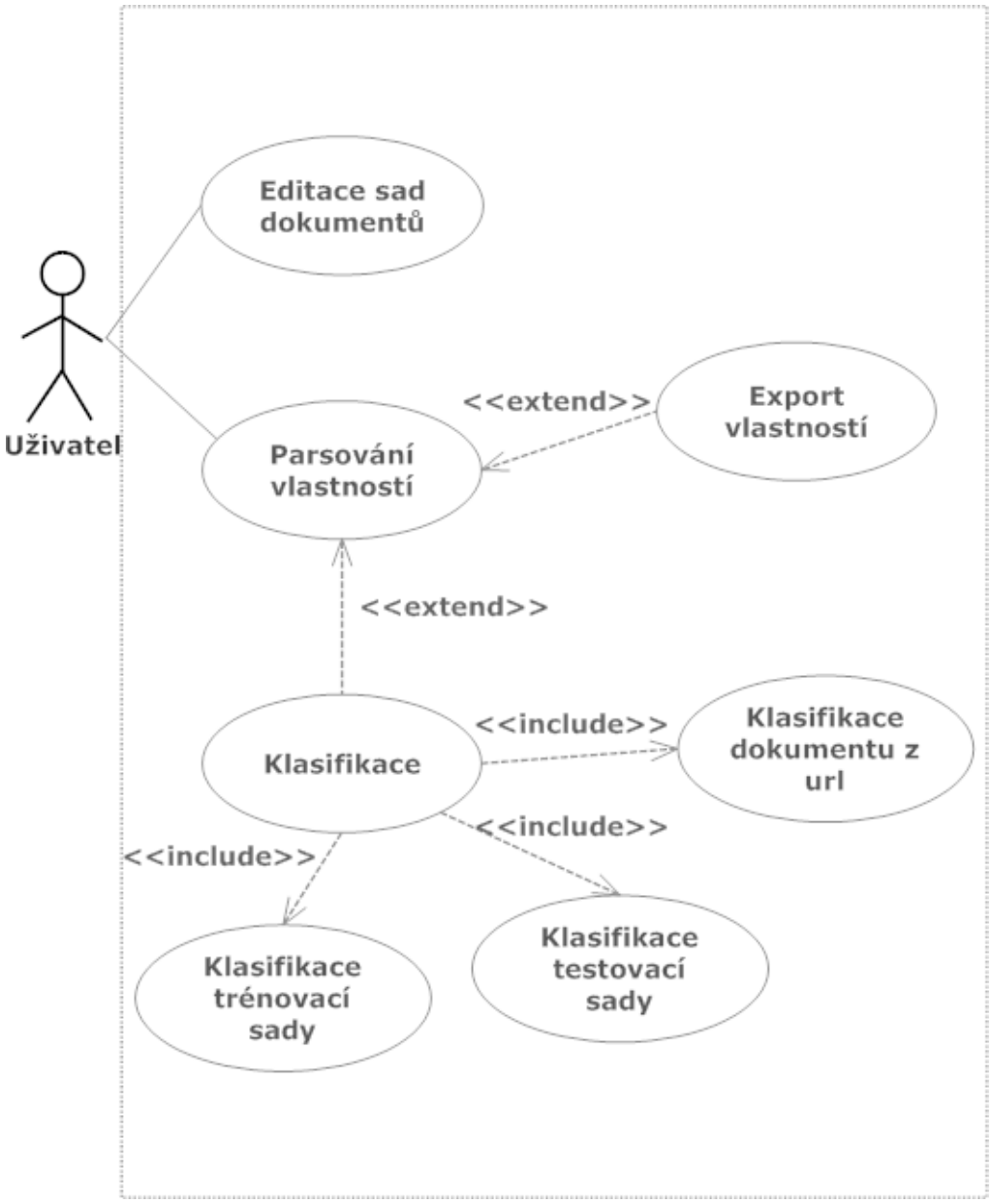
- learn - klasifikátor se naučí z vložených dat
- classify - klasifikátor vrátí název detekovaného žánru



Obrázek 2: Algoritmus klasifikace dokumentu pomocí Naivního bayesovského klasifikátoru

3.4 Aplikace

Pro otestování funkčnosti algoritmu jsem vytvořil aplikaci Web genre detector. Na obrázku 3 je zobrazen use case diagram popisující chování systému z pohledu uživatele.



Obrázek 3: Use case diagram aplikace

4 Experimenty

Všechny webové stránky použité v experimentech byly ručně shromážděny a přiřazeny do žánru. Při shromažďování stránek jsem se snažil vybírat příklady žánrů z různých domén a témat, aby byl pokryt co největší rozsah stránek.

Vyhodnocení výsledků

Abychom zjistili úspěšnosti testu, je potřeba výsledky analyzovat. V testování se bude počítat celková přesnost A_C (5), která nám určuje jak je test kvalitní [19].

$$A_C = (TP + TN)/(TP + TN + FP + FN) \tag{5}$$

TP znamená správně pozitivní (dokument je správně klasifikovaný jako žánr), TN znamená správně negativní (dokument je správně označen že nepatří do žánru), FP znamená falešně pozitivní (dokument je nesprávně klasifikovaný jako žánr) a FN znamená falešně negativní (dokument patřící do žánru nebyl klasifikován jako žánr).

4.1 Experiment - IF-THEN klasifikace

V prvním experimentu se testovalo zda webová stránka je žánr blog nebo není[4]. Pro tento experiment byly vlastnosti vybrány pozorováním.

Parametry testu

- trénovací sada - 20 dokumentů žánru blog, 20 dokumentů jiného žánru
- testovací sada - 20 dokumentů žánru blog, 20 dokumentů jiného žánru
- vlastnosti dokumentu - 2 funkčně motivované vlastnosti pro blog (viz. kapitola 2.4), typická slova pro žánr blog

[4]Tato metoda není implementována v aplikaci Web genre detector.

Výsledky

Ačkoli výsledky oproti pozdějším experimentů nejsou špatné, kvůli komplikované detekci funkčně motivovaných vlastností jsem od této metody upustil.

Žánr	Přesnost	TP	TN	FP	FN
BLOG	82.5 %	16	17	3	4

Tabulka 4: Úspěšnost IF-THEN klasifikace, trénovací sada

4.2 Popis experimentu - Naivní Bayesovský klasifikátor

V rámci testování úspěšnosti Naivního Bayesovského algoritmu jsem provedl 2 typy testů:

- Rozlišování zda dokument patří do zadaného žánru nebo nepatří - tzn. trénovací sada obsahuje dvě třídy (dokumenty patřící do žánru, dokumenty nepatřící do žánru).
- Přiřazení dokumentu do žánru ze známého počtu žánrů - tzn. trénovací sada obsahuje dokumenty ze 4 tříd (blog, diskuzní fórum, FAQ, elektronický obchod).

Výběr vhodných vlastností

Dobrá množina vlastností je ta, která obsahuje vlastnosti mající vztah s třídou, ale nemají vztah mezi sebou [16]. Pro náš experiment tedy bude důležité vybrat vhodné vlastnosti, podle nichž budeme moci klasifikovat dokument do žánru. Pro hledání vhodných vlastností jsem použil metodu Correlation-based Feature Subset Selection (CFS). K aplikování téty metody jsem použil program Weka[5]. Předchozích experimentech bylo dokázáno že vybráním vlastností pomocí CFS bylo dosaženo stejných nebo lepších výsledků u Naivního bayesovského klasifikátoru. CFS odstraní nerelevantní, nadbytečné a vlastnosti závislé na ostatních vlastnostech.

Postup experimentu

- Parser extrahuje všechny vlastnosti do csv souboru.

[5]Weka, open source program vyvinutý na universitě Waikato, je kolekce algoritmů strojového učení pro dolování dat (data mining). Weka obsahuje nástroje pro úpravu vstupních dat, klasifikaci, výběr atributů, vizualizaci atd[18].

- Vlastnosti se importují do programu Weka a aplikuje se metoda Correlation-based Feature Subset Selection pro výběr nejlepších vlastností.
- Klasifikátor se naučí z pouze vybraných vlastností.
- Klasifikátor klasifikuje všechny dokumenty v trénovací i testovací sadě.
- Na závěr bude vyhodnocení a porovnání výsledků s ostatními metodami strojového učení[6].

4.3 Experiment 1 - Naivní bayesovská metoda

V prvním experimentu se bude testovat zda stránka patří do žánru FAQ nebo nepatří.

Parametry testu

- trénovací sada - 20 dokumentů žánru FAQ, 20 dokumentů jiného žánru
- testovací sada - 20 dokumentů žánru FAQ, 20 dokumentů jiného žánru
- vlastnosti dokumentu - typické slova pro žánr FAQ, počet výskytů HTML tagu

Výsledky

Jak můžeme vidět v tabulkách 5 a 6 přesnost všech testovaných metod je stejná.

Žánr	Přesnost	TP	TN	FP	FN
FAQ	90 %	17	19	1	3

Tabulka 5: Úspěšnost klasifikace Naivní Bayes, testovací sada

FAQ	Naive Bayes	Bayes Net	LogitBoost	K-NN
Trénovací sada	100 %	100 %	100 %	100 %
Testovací sada	90 %	90 %	90 %	90 %

Tabulka 6: Úspěšnost klasifikace - porovnání metod

[6]Pro porovnání ostatních metod bude použit program Weka.

4.4 Experiment 2 - Naivní bayesovská metoda

V druhém experimentu se bude testovat zda stránka patří do žánru diskuze nebo nepatří.

Parametry testu

- trénovací sada - 20 dokumentů žánru diskuze, 20 dokumentů jiného žánru
- testovací sada - 20 dokumentů žánru diskuze, 20 dokumentů jiného žánru
- vlastnosti dokumentu - typické slova pro žánr diskuze

Výsledky

V druhém experimentu můžeme vidět o 2.5 % horší výsledky u naivní bayesovské metody.

Žánr	Přesnost	TP	TN	FP	FN
Diskuze	90 %	19	17	3	1

Tabulka 7: Úspěšnost klasifikace Naivní Bayes, testovací sada

FAQ	Naive Bayes	Bayes Net	LogitBoost	K-NN
Trénovací sada	95 %	97.5 %	97.5 %	97.5 %
Testovací sada	90 %	92.5 %	92.5 %	92.5 %

Tabulka 8: Úspěšnost klasifikace - porovnání metod

4.5 Experiment 3 - Naivní bayesovská metoda

V třetím experimentu se bude testovat zda stránka patří do žánru elektronický obchod nebo nepatří.

Parametry testu

- trénovací sada - 20 dokumentů žánru elektronický obchod, 20 dokumentů jiného žánru

- testovací sada - 20 dokumentů žánru elektronický obchod, 20 dokumentů jiného žánru
- vlastnosti dokumentu - typické slova pro žánr elektronický obchod, počet výskytů html tagu <img>

Výsledky

V třetím experimentu jsou vidět první větší rozdíly mezi metodami. Nejvyrovnanější výsledky má LogitBoost, u ostatních metod jsou znatelné rozdíly přesnosti mezi trénovací a testovací sadou.

Žánr	Přesnost	TP	TN	FP	FN
Elektronický obchod	90 %	16	20	0	4

Tabulka 9: Úspěšnost klasifikace Naivní Bayes, testovací sada

FAQ	Naive Bayes	Bayes Net	LogitBoost	K-NN
Trénovací sada	77.5 %	85 %	97.5 %	100 %
Testovací sada	90 %	80 %	95 %	85 %

Tabulka 10: Úspěšnost klasifikace - porovnání metod

4.6 Experiment 4 - Naivní bayesovská metoda

Ve čtvrtém experimentu se bude testovat zda stránka patří do žánru blog nebo nepatří.

Parametry testu

- trénovací sada - 20 dokumentů žánru blog, 20 dokumentů jiného žánru
- testovací sada - 20 dokumentů žánru blog, 20 dokumentů jiného žánru
- vlastnosti dokumentu - typické slova pro žánr blog, počet výskytů nejčastějšího nadpisu, počet znaků

Výsledky

U detekce blogu je vidět velký rozdíl v přesnosti na trénovací sadě u Naivního bayesovské klasifikátoru oproti ostatním metodám. Na testovací sadě již tolik nezaostává.

Žánr	Přesnost	TP	TN	FP	FN
Blog	67.5 %	10	17	3	10

Tabulka 11: Úspěšnost klasifikace Naivní Bayes, testovací sada

FAQ	Naive Bayes	Bayes Net	LogitBoost	K-NN
Trénovací sada	72.5 %	92.5 %	100 %	100 %
Testovací sada	67.5 %	67.5 %	75 %	75 %

Tabulka 12: Úspěšnost klasifikace - porovnání metod

4.7 Experiment 5 - Naivní bayesovská metoda

Parametry testu

- trénovací sada - 20 dokumentů žánru blog, elektronický obchod, FAQ a diskuzní fórum (celkem 80 dokumentů)
- testovací sada - 20 dokumentů žánru blog, elektronický obchod, FAQ a diskuzní fórum (celkem 80 dokumentů)
- vlastnosti dokumentu - typické slova pro jednotlivé žánry, průměrná délka věty, 50 nejběžnějších slov v angličtině, počet výskytů html tagu <img>

Výsledky

V posledním testu se rozlišovalo mezi čtyřmi žánry, nejhůře dopadla detekce blogu. A to pravděpodobně z důvodu, že jednotlivé testované dokumenty tohoto žánru se navzájem nejvíce lišily. Nejlepších výsledků dosahovala metoda LogitBoost.

Žánr	Přesnost	TP	TN	FP	FN
Blog	47.5 %	11	8	12	9
Elektronický obchod	67.5 %	14	13	7	6
Diskuze	77.5 %	16	15	5	4
FAQ	72.5 %	12	17	3	8

Tabulka 13: Úspěšnost klasifikace Naivní Bayes, testovací sada

FAQ	Naive Bayes	Bayes Net	LogitBoost	K-NN
Trénovací sada	86.25 %	92.5 %	100 %	100 %
Testovací sada	66.25 %	76.25 %	80 %	72.5 %

Tabulka 14: Úspěšnost klasifikace - porovnání metod

4.8 Shrnutí výsledků experimentů

Jak můžete vidět v tabulce 15 nejlepších výsledků ve všech testech dosáhla metoda LogitBoost. Pro praktické využití jsou neužitečnější výsledky testů na testovací sadě. Zde Naivní bayesovská metoda zaostávala nejméně oproti ostatním metodám.

FAQ	Naive Bayes	Bayes Net	LogitBoost	K-NN
Trénovací sada	86.25 %	93.5 %	99 %	99.5 %
Testovací sada	80.75 %	81.25 %	86.5 %	83 %

Tabulka 15: Průměrná úspěšnost klasifikace všech testů - porovnání metod

5 Závěr

Cílem práce bylo navrhnout aplikaci a metodu pro detekci žánrů webových stránek. Výsledná aplikace Web genre detector používá k detekci žánrů pravděpodobnostní Naivní bayesovskou metodu.

V mé práci jsem testoval 4 webové žánry - blog, diskuzní fórum, FAQ a elektronický obchod. Jelikož se aplikace učí z trénovacích dat, může při použití jiných extrahovaných vlastností detekovat téměř jakýkoliv žánr. Aplikace měla největší úspěšnost u strukturálně jednodušších žánrů jako FAQ a diskuzní fórum. U elektronického obchodu a u blogů byly pozorovány horší výsledky, a to hlavně z důvodu rozdílné struktury a obsahu testovaných příkladů. V porovnání s ostatními metodami Naivní bayesovský klasifikátor mírně zaostává a pro další práce by bylo lepší využít sofistikovanější algoritmus (např. LogitBoost).

Největší slabinu této metody spatřuji v parsování dokumentu, což by se v budoucích pracích jistě mělo zlepšit. Metoda používá jednoduché vlastnosti (počty HTML tagů, frekvence výrazů) namísto složitějších funkčně motivovaných vlastností, které lépe charakterizují žánr. Bohužel detekovat tyto vlastnosti počítačem není snadné. Domnívám se pokud bychom použili stejnou klasifikační metodu a místo vlastností detekované parserem použili vlastnosti detekované člověkem, výsledky by byli několikanásobně lepší. Svou vinu na horších výsledcích nese také otevřenost internetu, kde neexistuje žádná šablona jak by měl žánr vypadat. Mírné zlepšení detekce žánrů vidím v HTML 5, která na rozdíl od HTML 4 má sémantické prvky pro přesnější vyznačení částí dokumentu.

6 Reference

[1] BOESE, Elizabeth. *STEREOTYPING THE WEB: GENRE CLASSIFICATION OF WEB DOCUMENTS* [online]. 2005 [cit. 2012-04-16].
Dostupné z: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.112.3660&rep=rep1&type=pdf

[2] ROSSO, M. A. *User-based identification of Web genres* [online]. 2008 [cit. 2012-04-19]. Dostupné z: http://onlinelibrary.wiley.com/doi/10.1002/asi.20798/full

[3] KEH-YIH SU, JUN'ICHI TSUJII, JONG-HYEOK LEE a OI YEE KWONG. *Automatic Genre Detection of Web Documents*. [online]. 2004[cit. 2012-04-02]. Dostupné z: http://www.springerlink.com/content/ra44fax56qmdrk0w/

[4] SHAROFF, S., A. MEHLER a M. SANTINI. *Genres on the web*: *computational models and empirical studies*. Dordrecht: Springer, 2010. ISBN 90-481-9177-7.

[5] HONZÍK, Petr. *Strojové učení*. [online]. 2006[cit. 2012-04-03]. Dostupné z: http://www.umel.feec.vutbr.cz/VIT/images/pdf/studijni_materialy/ing/Strojove_uceni_S.pdf

[6] BOESE, Elizabeth Sugar. *Stereotyping the web: genre classification of web documents*. [online]. 2005[cit. 2012-04-03]. Dostupné z: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.112.3660&rep=rep1&type=pdf

[7] N. DEWDNEY, C. VanEss-Dykema, and R. MacMillan. *The form is the substance: classification of genres in text*. [online]. 2001[cit. 2012-04-03]. Dostupné z: http://dl.acm.org/citation.cfm?id=1118227

[8] SERRANO, J. I., M. TOMEČKOVÁ a J. ZVÁROVÁ. *Metody strojového učení pro vyhledávání znalostí v lékařských datech o ateroskleróze* [online]. [cit. 2012-04-16]. Dostupné z: http://www.ejbi.org/articles/200608/25/2.html

[9] Mehler, Rehm, Santini, Sharoff. *WebGenreWiki* [online]. [cit. 2012-03-28]. Dostupné z: http://www.webgenrewiki.org/index.php5/Definition

[10] MÉZL, Martin. *BAYESIAN METHODS FOR DATA MINING* [online]. [cit. 2012-04-07]. Dostupné z: http://www.feec.vutbr.cz/EEICT/2009/sbornik/02-Magisterske%20projekty/10-Inteligentni%20systemy/05-xmezlm00.pdf. Diplomová práce. VUT.

[11] Naive Bayes classifier. In: *Wikipedia: the free encyclopedia* [online]. San Francisco (CA): Wikimedia Foundation, 2001- [cit. 2012-03-30]. Dostupné z: http://en.wikipedia.org/wiki/Naive_Bayes_classifier

[12] BARTONČÍK, MICHAL. *ROZPOZNÁVÁNÍ VÝRAZU TVÁŘE U NEZNÁMÝCH OSOB* [online]. 2011 [cit. 2012-04-21]. Dostupné z: http://www.vutbr.cz/www_base/zav_prace_soubor_verejne.php?file_id=38405

[13] Support vector machine. In: *Wikipedia: the free encyclopedia* [online]. San Francisco (CA): Wikimedia Foundation, 2001- [cit. 2012-04-12]. Dostupné z: http://en.wikipedia.org/wiki/Support_vector_machine

[14] Blog. In: *Wikipedia: the free encyclopedia* [online]. San Francisco (CA): Wikimedia Foundation, 2001- [cit. 2012-04-1]. Dostupné z: http://en.wikipedia.org/wiki/Blog

[15] FAQ. In: *Wikipedia: the free encyclopedia* [online]. San Francisco (CA): Wikimedia Foundation, 2001- [cit. 2012-04-1]. Dostupné z: http://en.wikipedia.org/wiki/FAQ

[16] HALL, Mark A. *Correlation-based Feature Selection for Machine Learning* [online]. 1999 [cit. 2012-04-13]. Dostupné z: http://www.lri.fr/~pierres/donn%E9es/save/these/articles/lpr-queue/hall99correlationbased.pdf

[17] SANTINI, Marina. *Description of 3 feature sets for automatic identification of genres in web pages* [online]. 2005 [cit. 2012-04-13]. Dostupné z: http://www.nltg.brighton.ac.uk/home/Marina.Santini/three_feature_sets.pdfforMachineLearning

[18] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, Ian H. Witten (2009); The WEKA Data Mining Software: An Update; SIGKDD Explorations, Volume 11, Issue 1.

[19] KAŠPAROVÁ, M., J. KŘUPKA a J. PÍRKO. *MODELOVÁNÍ SPOKOJENOSTI OBČANŮ VE VZTAHU K REGIONÁLNÍMU ROZVOJI A KVALITĚ ŽIVOTA* [online]. 2008 [cit. 2012-04-19]. Dostupné z: http://dspace.upce.cz/xmlui/bitstream/handle/10195/35087/KasparovaM_Modelovani%20spokojenosti_SP%20FES_2008.pdf?sequence=1

A Příloha na CD/DVD

Součástí práce je CD s elektronickou podobou práce ve formátu PDF. Dále jsou zde umístěny testované dokumenty a zdrojové kódy.

/corpus Testované dokumenty

/source Zdrojové kódy aplikace

/text Elektronická podoba práce

B Zdrojové kódy

Na výpisech 5 a 6 jsou vybrané ukázky zdrojových kódů použitých při parsování dokumentu.

```
//odstrani html tagy, scripty, komentare, prazdne znaky
private string ConvertToText(string html)
{
    //odstraneni scriptu
    html = Regex.Replace(html, "<script.*?</script>", string.Empty, RegexOptions.
        Singleline | RegexOptions.IgnoreCase);
    //odstraneni komentaru
    html = Regex.Replace(html, "<!--.*?-->", string.Empty, RegexOptions.Singleline |
        RegexOptions.IgnoreCase);
    //odstraneni html tagu
    html = Regex.Replace(html, "<[^>]*>", string.Empty);

    //nahrazeni bilych znaku
    html = Regex.Replace(html, "\n", " ");
    html = Regex.Replace(html, "\r", " ");
    html = Regex.Replace(html, "\t", " ");

    //nahrazeni html entit
    html = Regex.Replace(html, " ", " ");
    html = Regex.Replace(html, "&lt;", "<");
    html = Regex.Replace(html, "&gt;", ">");
    html = Regex.Replace(html, "&amp;", "&");
    html = Regex.Replace(html, "&quot;", "\"");
    html = Regex.Replace(html, "&plusmn;", "$\pm$");
    html = Regex.Replace(html, "&times;", "$\times$");
    html = Regex.Replace(html, "&reg;", "$\textregistered$");
    html = Regex.Replace(html, "&copy;", "$\copyright$");
    html = Regex.Replace(html, "&euro;", "$\texteuro$");
    html = Regex.Replace(html, "&raquo;", "$\guillemotright$");
    html = Regex.Replace(html, "&laquo;", "$\guillemotleft$");

    //odstraneni opakovanych mezer
    html = Regex.Replace(html, "( )+", " ");
    return html;
}
```

Výpis 5: Převede HTML vstup na čistý text

```
//prumerna delka vety
HtmlNodeCollection nodes = document.DocumentNode.SelectNodes("//p");
int senCount = 1;
long totalSenLen = 0;
if (nodes != null)
{
    foreach (HtmlNode n in nodes)
    {
        string s = this.ConvertToText(n.InnerText);
        //pridani mezery na konec, kvuli detekci vet
        s += " ";
        matchCol = Regex.Matches(s, @"(?sx-m)[^\r\n].*?(?:(?:\.|\?|!)\s)",
            RegexOptions.Compiled | RegexOptions.IgnoreCase);
        foreach (Match m in matchCol)
        {
            totalSenLen += m.Length;
            senCount++;
        }
    }
    parsedDoc.AvgSenLen = totalSenLen / senCount;
}
else
{
    parsedDoc.AvgSenLen = 0;
}
```

Výpis 6: Příklad spočítaní průměrné délky věty

C Uživatelská příručka

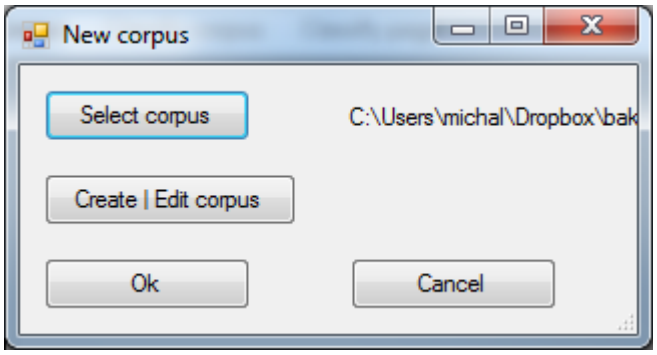
Pro ověření funkčnosti implementované metody jsem vytvořil aplikaci Web genre detector. Aplikace je naprogramována v jazyce C# s využitím .NET frameworku. V uživatelské příručce bude popsáno ovládání aplikace a její možnosti. Po spuštění aplikace se zobrazí prázdné okno, základním prvkem uživatelského rozhraní je menu s nabídkami:

- Corpus - zde jsou umístěny prvky pro práci s trénovací sadou, parsování dokumentů, export vlastností dokumentu.
- Classify corpus - tato nabítka slouží ke klasifikaci vybrané sady uložené na disku.
- Classify page - po vložení url aplikace klasifikuje webovou stránku.

Práce s sadou dokumentů (Nabídka Corpus)

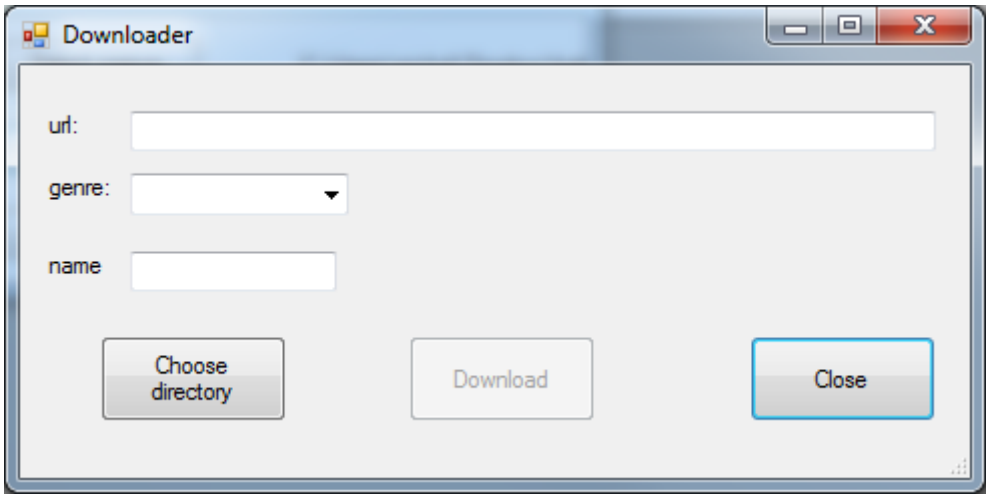
Obsahuje několik podnabídek:

- Select - zobrazí se dialogové okno (viz. obrázek 4), kde můžeme vybrat existující sadu (Select corpus) nebo vytvořit novou sadu (Create | Edit corpus). Trénovací sada je adresář obsahující upravené HTML dokumenty s informací o žánru a url dokumentu.
- Learn - zobrazí se dialogové okno (viz. obrázek 6) pro výběr vlasností, které má parser extrahovat. Po zparsování dokumentu se klasifikátor naučí a získaná data se zobrazí v tabulce.
- Export features - exportuje vlastnosti dokumentu do csv souboru.

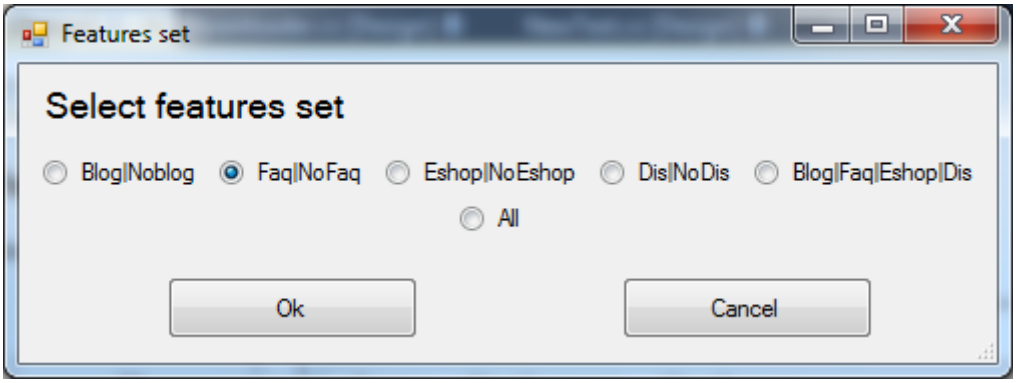


Obrázek 4: Výběr sady dokumentů

Po kliknutí na tlačítko Create | Edit corpus se zobrazí okno Downloader (viz. obrázek 5). Po zadání url, žánru a názvu se dokument se stáhne do určeného adresáře.



Obrázek 5: Stažení a přiřazení HTML dokumentů do žánru



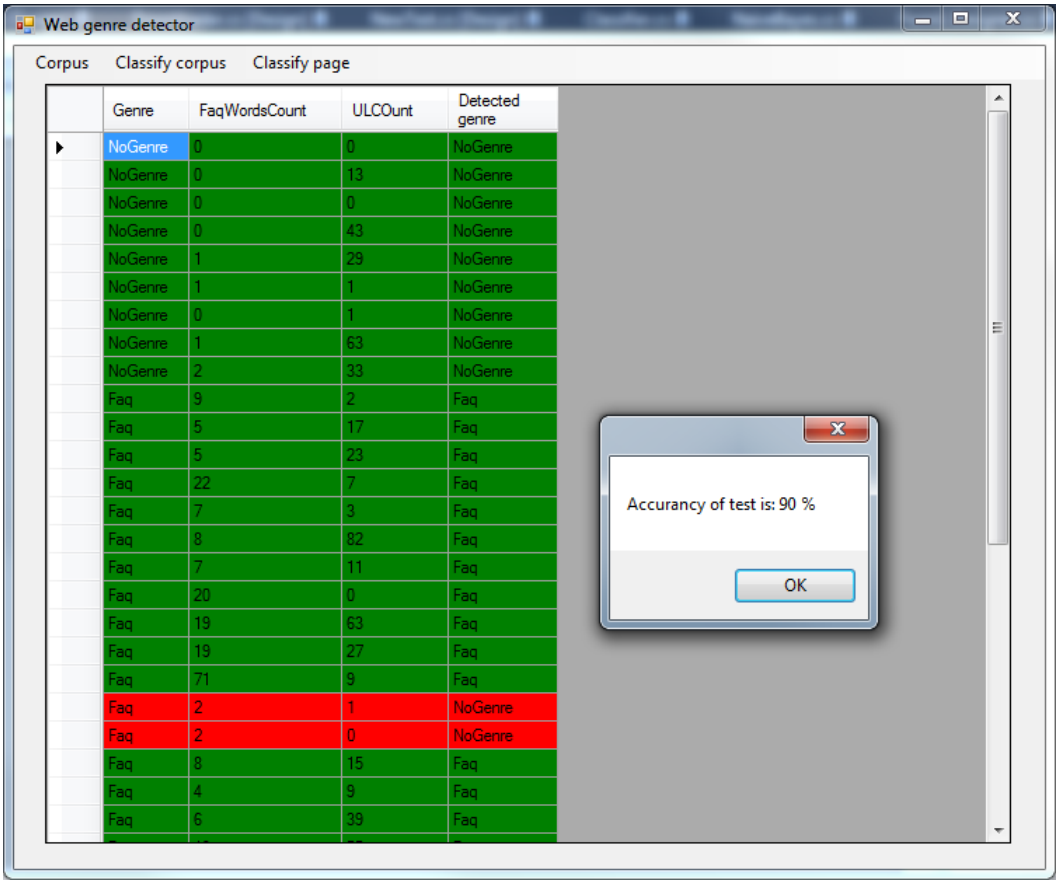
Obrázek 6: Výběr vlastností, které se budou extrahovat

Klasifikace sady dokumentů (Nabídka Classify corpus)

Obsahuje dvě podnabídky:

- Use training set - klasifikovat se bude trénovací sada.
- Use test set - klasifikovat se bude testovací sada.

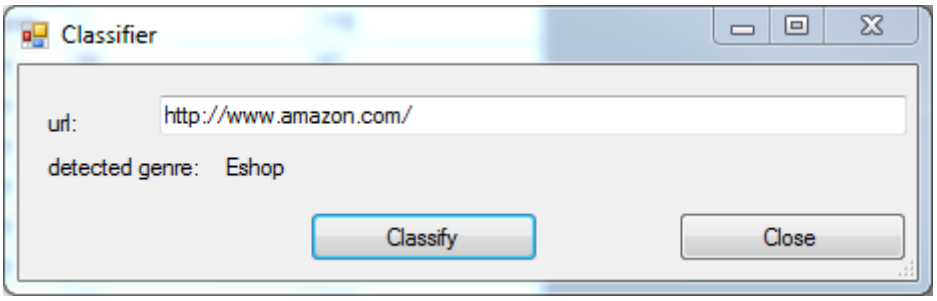
Po kliknutí na jednu z podnabídek bude každý dokument klasifikován, zobrazí se procentuální úspěšnost testu. Správně klasifikované dokumenty jsou označeny zeleně, nesprávně klasifikované dokumenty jsou označeny červeně.



Obrázek 7: Výběr vlastností, které se budou extrahovat

Klasifikace internetové stránky (Nabídka Classify page)

Poslední nabídka nám umožňuje klasifikovat webovou stránku do jednoho z naučených žánrů (viz obrázek 8).



Obrázek 8: Výběr sady dokumentů